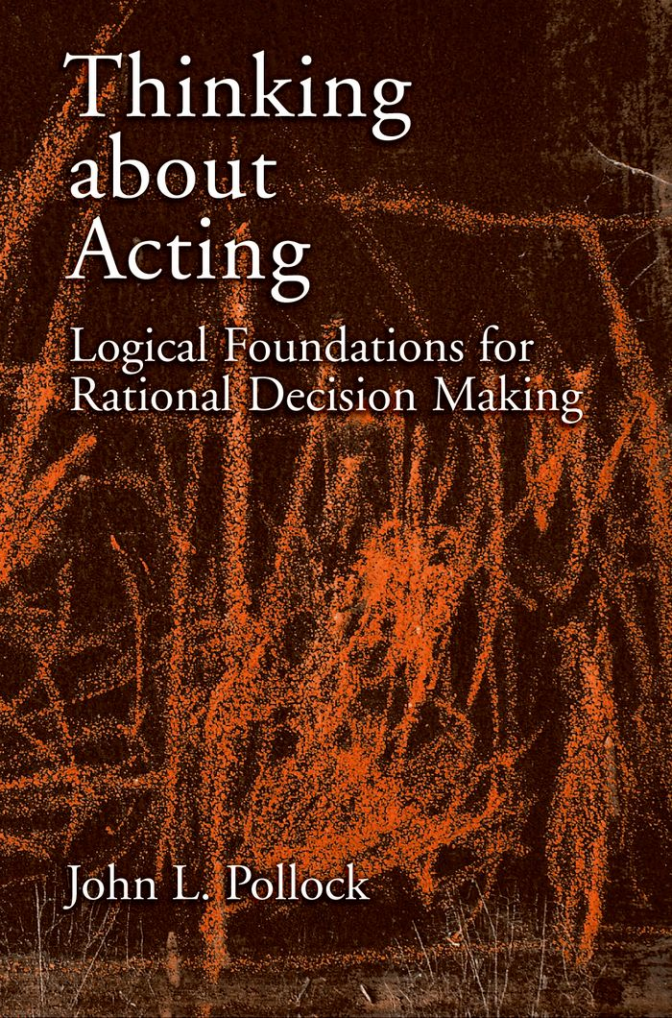
Thinking about Acting

Logical Foundations for Rational Decision Making



John L. Pollock

Thinking about Acting

This page intentionally left blank

Thinking about Acting

Logical Foundations for Rational Decision Making

John L. Pollock

OXFORD
UNIVERSITY PRESS

2006

OXFORD
UNIVERSITY PRESS

Oxford University Press, Inc., publishes works that further
Oxford University's objective of excellence
in research, scholarship, and education.

Oxford New York
Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in
Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Published by Oxford University Press, Inc.
198 Madison Avenue, New York, New York 10016

www.oup.com

Oxford is a registered trademark of Oxford University Press

Library of Congress Cataloging-in-Publication Data
Pollock, John L.
Thinking about acting : logical foundations for rational
decision making / John L. Pollock.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-19-530481-7
1. Statistical decision. 2. Decidability (Mathematical logic)
3. Probabilities. 4. Induction (Logic) I. Title.
QA279.4.P65 2006
519.5'42—dc22 2005054921

Printed in the United States of America
on acid-free paper

For Lilian

This page intentionally left blank

Preface

The objective of this book is to produce a theory of rational decision making for realistically resource-bounded agents. My interest is not in "What should I do if I were an ideal agent?" but rather, "What should I do given that I am who I am, with all my actual cognitive limitations?"

The book has three parts. Part I addresses the source of the values that agents use in rational decision making. The most common view among philosophers and cognitive scientists is that the primitive evaluative database that real agents employ in evaluating outcomes is a preference ranking, but I argue that this is computationally impossible. An agent's evaluative database must instead assign real numbers to outcomes. I argue that, contrary to initial appearances, this is psychologically plausible.

Part II investigates the knowledge of probability that is required for decision-theoretic reasoning. I argue that subjective probability makes no sense as applied to real (resource bounded) agents. Rational decision making must instead be based on a species of objective probability. Part II goes on to sketch a theory of objective probability. I use that to define a variety of causal probability and argue that this is the kind of probability presupposed by rational decision making.

Part III explores how these values and probabilities are to be used in decision making. Classical decision theory is based on the optimality principle, according to which rationality dictates choosing actions that constitute optimal solutions to practical problems. Optimality is defined in terms of expected values. I will argue that the optimality prescription is wrong, for several reasons: (a) actions cannot be chosen in isolation—they must be chosen as parts of plans; (b) we cannot expect real agents to find optimal plans, because there are infinitely many alternatives to survey; (c) plans cannot be evaluated in terms of their expected values anyway, because different plans can be of different scopes. I construct an alternative, called "locally global planning", that accommodates these difficulties. According to locally global planning, individual plans are to be assessed in terms of their contribution to the cognizer's "master plan". Again, the objective cannot be to find master plans with maximal expected values, because there may be none, and even if there are any, finding them is not a computationally feasible task for real agents. Instead, the objective must be to find good master plans, and improve them as better ones come along. It is argued that there are computationally feasible ways of doing this, based on defeasible reasoning about values and probabilities.

This work is part of the OSCAR project, whose objective is to construct an implementable theory of rational cognition and implement it in an AI system. This book stops short of implementation, but that is the next step. This book provides the theoretical foundations for an implemented system of decision-theoretic planning, and future research will push the work in

that direction.

Much of the material presented in this book has been published, in preliminary form, in other places, and I thank the publishers of that material for allowing it to be reprinted here. Much of Part I is drawn from "Evaluative cognition" (*Nous*, **35**, 325–364). Chapter 8 is based upon "Causal probability" (*Synthese* **132**, 143–185). Chapter 9 is based upon "Rational choice and action omnipotence" (*Philosophical Review* **111**, 1–23). Chapter 10 and part of chapter 12 are based upon "Plans and decisions" (*Theory and Decision* **57**, 79–107) and "Against optimality: Logical foundations for decision-theoretic planning in autonomous agents" (*Computational Intelligence* **22**). The appendix is a revised version of "The theory of nomic probability" (*Synthese* **90**, 263–300).

I also thank the University of Arizona for its support of my research. I particularly want to thank Merrill Garrett for his continued enthusiasm for my work and the help he provided in his role as Director of Cognitive Science, and I want to thank Chris Maloney for his steadfast support as Head of the Department of Philosophy. I am indebted to numerous graduate students for their unstinting constructive criticism, and to my colleagues for their interactions over the years. I want to mention specifically Douglas Campbell, Josh Cowley, Justin Fisher, and Nicole Hassoun, who helped me more than I can say.

This work has been supported by grants no. IRI-9634106 and IRI-0080888 from the National Science Foundation.

Contents

Part III: Decisions

This page intentionally left blank

Thinking about Acting

This page intentionally left blank

1
Rational Choice and Classical Decision Theory

1. Rational Cognition

We make decisions constantly, at almost every moment of our waking lives. Most are little decisions—"Should I put more mustard on my sandwich?" But some are momentous—"Should I marry Jane?" Some people are better decision makers than others, and some decisions are better than others. What makes one decision better than another? One sense in which a decision can be better is that it has a better outcome. But there is also an internal dimension of criticism. A decision can be evaluated as having been made well or badly regardless of its outcome. Because Claudio was furious with Flavia, he spent his paycheck on lottery tickets rather than paying the mortgage. He got lucky and won, and they are now millionaires, but it was still a stupid thing to do. His decision was irrational.

What makes a decision rational or irrational? How should we go about making decisions so that they are rational? That is the topic of this book. I want to know how we, as human beings, should go about deciding what actions to perform.

We are cognitive agents. Cognitive agents think about the world, evaluate various aspects of it, reflect upon how they might make it more to their liking, and act accordingly. Then the cycle repeats. This is the *doxast-conative* loop, diagrammed in figure 1.1. The defining characteristic of cognitive agents is that they implement the doxastic-conative loop by thinking about the world and acting upon it in response to their deliberations. Both human beings and the autonomous rational agents envisaged in AI are cognitive agents in this sense.

This cognition can be divided roughly into two parts. *Epistemic cognition* is that kind of cognition responsible for producing and maintaining beliefs. *Practical cognition* evaluates the world, adopts plans, and initiates action. We can further divide practical cognition into three parts: (1) the evaluation of the world as represented by the agent's beliefs, (2) the selection of actions or plans aimed at changing it, and (3) the execution of the plans.

Some aspects of our cognition are beyond our control, and it makes no sense to ask how we should perform those cognitive tasks. For example, when I look at the world, purely automatic computational processes take as

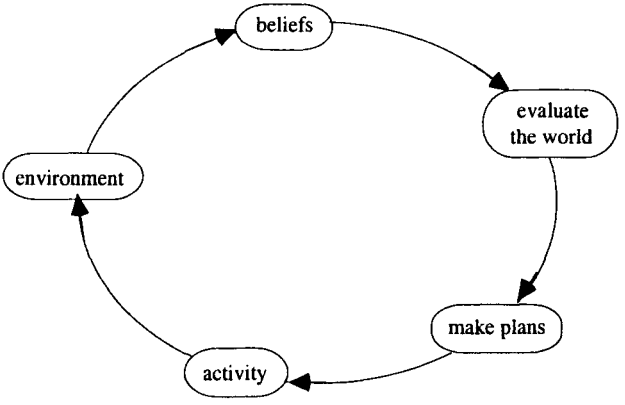


Figure 1.1 The doxastic-conative loop

input the pattern of stimulation at my optic nerve and produce a visual image. The visual image provides my visual access to the world. But I have no control over how the image is produced. If I see a newspaper illuminated by what I know to be red light, and the newspaper looks red to me, I cannot be criticized as irrational because it looks red to me. That is beyond my control. But I can be criticized as irrational if I believe on the basis of the visual image that the newspaper really is red. This is because the inference is something over which I have a certain amount of control. I can at the very least withdraw my conclusion in light of my knowledge that newspapers are generally white and my knowledge of how red lights can make white things look red. But there is nothing I can do to make the newspaper stop looking red to me.

A theory of rationality is a theory about how a cognitive agent should perform the kinds of cognitive tasks over which it has some control.[1] Just as cognition divides roughly into epistemic cognition and practical cognition, so rationality divides roughly into epistemic rationality and practical rationality. Epistemology studies epistemic rationality, and I have written about that extensively elsewhere.[2] The focus of this book is practical rationality. I want to know how a cognitive agent should go about deciding what actions to perform. An answer to this question constitutes a *theory of rational choice.* So this is a book about rational choice.

My principal concern is with human decision making. I want to know how we, as human beings, should decide what actions to perform. However, idiosyncratic features of human psychology sometimes obscure the logic of

[1] This point is developed more fully in Pollock (2006).

[2] See particularly my (1986, 1995) and Pollock and Cruz (1999).

rational decision making, and we can often clarify the issues by focusing more broadly on rational decision making in any cognitive agent, human or otherwise. Humans are the most sophisticated cognizers we currently know about, but we can usefully ask how *any* cognitive agent should go about deciding how to act. The results of this investigation should be as applicable to the construction of artificial rational agents in AI as to human beings. The advantage of taking this broader perspective is that it can sometimes be argued that purely computational considerations illuminate issues in the theory of rational choice, showing that theories motivated by thinking specifically about human beings cannot be correct for any cognitive agents, and so in particular they cannot be correct for human beings.

The term "practical reasoning" has been used ambiguously in philosophy, on the one hand to refer to purely self-interested reasoning about action, and on the other hand to include the moral aspects of decision making. As I use the terms "practical reasoning" and "practical cognition" in this book, they are about purely self-interested decision making. An individual comes to a decision problem with various goals and then tries to select actions that will achieve those goals. I want to know how such decisions should be made. The "should" here is a practical "should", not a moral "should". The problems of morality are orthogonal to understanding practical cognition in this sense. Morality could interact with practical cognition in various ways. It *might* function by simply adding goals to be achieved by practical cognition, or by affecting the evaluation of goals. In either case, morality would function via the mechanisms of practical cognition, and would not be in conflict with it. But morality might also function in a way that puts it in potential conflict with self-interested practical decision making. Moral philosophers have endorsed both of these views of the relationship between morality and practical decision making. In this book, however, I propose to remain neutral on issues of morality.

2. Ideal Rationality and Real Rationality

Human beings, and any real cognitive agents, are subject to cognitive resource constraints. They have limited reasoning power, in the form of limited computational capacity and limited computational speed. This makes it impossible, for example, for them to survey all of the logical consequences of their beliefs, or to compare infinitely many alternatives. This is a fundamental computational fact about real agents in the real world, and I would suppose that it could not have been otherwise. An account of how a real agent should make decisions must take account of these limitations.

Theories of rational action are sometimes taken to be theories about how ideal agents, immune to such cognitive limitations, should make decisions (Cherniak 1986; Skyrms 1980, 1984; Lewis 1981). One can, of course, choose to talk that way, but it is hard to see what that has to do with what we, as fallible human beings, should do. For instance, if a theory of ideal agents says that they should attend to all of the logical consequences of their

beliefs, but we as human beings cannot do that, then the recommendations applicable to ideal agents are simply not applicable to us. *We* should do something else. As I use the term "theory of rational action", it is about what we, and other resource bounded cognitive agents, should do. I want to know how, given our cognitive limitations, we should decide what actions to perform. In other words, I want a theory of *real rationality* as opposed to a theory of *ideal rationality*.

This distinction is widely recognized, but it often seems to be supposed that as philosophers our interest should be in ideal rationality. The rationality a human can achieve is mere "bounded rationality"—a crude approximation to ideal rationality. But surely we come to the study of rational decision making with an initial interest in how we, and agents like us, should make decisions. This is the notion of rationality that first interests us, and this is what I am calling "real rationality". We might try to illuminate real rationality by taking it to be some kind of approximation to ideal rationality, but still our original interest is in real rationality.

Although theories of ideal agents are not directly about how real agents should solve decision problems, a plausible suggestion is that the rules of rationality for real agents should be such that, as we increase the reasoning power of a real agent, insofar as it behaves rationally its behavior will approach that of an ideal rational agent in the limit. This is to take theories of ideal rationality to impose a constraint on theories of real rationality. We can make this suggestion more precise by distinguishing, as I have elsewhere (1986, 1995), between "justified" choices and "warranted" choices. A *justified* choice is one that a real agent could make given all of the reasoning it has performed *up to the present time* and without violating the constraints of rationality. A *warranted* choice is one that would be justified if the agent could complete all possibly relevant reasoning. Two characteristics of real agents make this distinction important. First, for any cognitively sophisticated agent, reasoning is non-terminating. There will never be a point at which the agent has completed all the reasoning that could possibly be relevant to a decision. But agents have to act. They cannot wait for the completion of a non-terminating process, so decisions must be made on the basis of the reasoning that has been done so far. In other words, real agents must act on the basis of justified choices rather than waiting until they know that a choice is warranted. Second, it is characteristic of the reasoning of a real agent that almost all of its conclusions are drawn *defeasibly*. That is, the reasoning to date can make the conclusion justified, but acquiring additional information or performing additional reasoning may rationally necessitate the agent's changing its mind.[3]

For an agent that reasons defeasibly, we can characterize a warranted choice as one that, at some stage of its reasoning, the agent could settle on and never subsequently have to change its mind no matter how much

[3] For the most part, it will be unimportant in this book exactly how defeasible reasoning works. I have, however, discussed it at length elsewhere. See my (1995, 2002), and Pollock and Cruz (1999).

additional reasoning it might perform. This can be made more precise by talking about "stages of reasoning". The agent starts from some initial epistemic situation, and then at each stage of reasoning it either draws a new conclusion or retracts a previous conclusion. A conclusion (or choice) is warranted iff there is a stage such that (1) it is justified at that stage, and (2) it remains justified at all subsequent stages of reasoning.[4]

The warranted choices are those an ideal agent that was able to perform all relevant reasoning would make on the basis of the information currently at its disposal. One might suppose that warranted choices are those we want an agent to make. The difficulty is that a real agent cannot complete all the reasoning that might possibly be relevant to a decision. As remarked above, reasoning is a non-terminating process. Eventually the agent has to act, so we cannot require that it act only on the basis of warranted choices. The most we can require is that the agent perform a "respectable amount" of reasoning, and then base its choice on that. So a real agent acts on the basis of justified choices that might not be warranted.

In some cases it would actually be irrational for a real agent to make the warranted choice. For instance, suppose P and Q are logically equivalent, but the agent has not yet performed enough reasoning to know this. Suppose the agent has good reason to accept a bet that P is true at 2:1 odds. Suppose that choice is not only justified, but also warranted. Suppose, however, the agent has no basis for assessing the probability of Q. That is, it has no justified beliefs about the probability of Q. Then it would be irrational for the agent to accept a bet that Q is true at 2:1 odds. That choice would not be justified. But it would be the warranted choice, because if the agent performed *enough* reasoning it would discover that Q is equivalent to P and hence has the same probability.

Theories of ideal agents are theories of warrant. It might be suggested that the behavior of an ideal agent is the target at which real agents should aim, and hence theories of real rationality can be evaluated in terms of whether they approach the correct theory of ideal rationality in the limit. More precisely, a theory of real rationality, viewed as a theory of justified choice, implies a theory of warrant. We can think of a theory of ideal rationality as a theory of what the correct theory of warrant should say. The suggestion would then be that a theory of justified choice (real rationality) is correct iff its implied theory of warrant describes the behavior of an ideal rational agent (given some theory of what ideal rationality requires).

For epistemic cognition, real rationality and ideal rationality might be related in some such fashion, but it will turn out that there can be no such connection in the case of practical cognition. The set of justified choices will only converge to the set of warranted choices if there are always warranted choices to be made. But it will emerge in chapter 10 that there may often be no warranted choices for real agents living in the real world. It could be that no matter how good a solution the agent finds for a decision problem,

---

[4] There are two different concepts of warrant here. For a discussion of their interconnections, see chapter 3 of my (1995).

given enough time to reason there is always a better solution to be found. I will argue that this need not be an untoward result. The supposition that there must always be warranted choices turns on a misunderstanding of the logical structure of practical cognition—it assumes that decision problems always have optimal solutions. If they do not, then theories of warrant would seem to be irrelevant to theories of justified decision making.

So our target is a theory of real rationality—a theory of how real agents, with all their cognitive limitations, should make decisions about how to act. A theory of ideal rationality might conceivably be relevant to the construction of such a theory, somehow imposing constraints on it, but a theory of ideal rationality by itself cannot solve the problem of producing a theory of real rationality.

3. Human Rationality and Generic Rationality

A theory of real rationality is a theory of how one should proceed in making decisions. We might put this by saying that our concept of rationality is a procedural concept. I have discussed procedural rationality at length elsewhere in connection with epistemic cognition.[5] An agent's cognitive architecture determines how the agent goes about performing various tasks. However, as remarked in section 1, the human cognitive architecture leaves us some leeway in how to perform many tasks. Various cognitive tasks are under our control to some degree, and a theory of rationality aims at telling us how we should perform those tasks.

It is the fact that we have control over our own cognition that makes it possible for us to behave irrationally. When we can choose how to perform a cognitive task, we can do the wrong thing, thereby proceeding irrationally. So, for example, we conclude that agents should not engage in wishful thinking or hasty generalization, but observe that, nevertheless, they sometimes do. It is interesting to inquire why humans are so constructed that it is possible for them to behave irrationally. Why aren't we built so that it is rigidly determined that we always do the right thing? Sometimes this is because having the power to control the course of our thinking makes us more efficient problem solvers. But that same power enables us to behave irrationally. For instance, one thing we have control over is what to think about. By enabling a cognitive agent to engage in practical cognition about what to think about we enable the agent to focus on problems that it is more apt to be able to solve and to try to solve them in ways it thinks are more likely to be successful. But this same power to control what it thinks about enables an agent to *avoid* thinking about something. In particular, if the agent has a favored theory but has reason for suspecting that some particular consideration may constitute a problem for the theory, the agent

[5] I introduced the concept of procedural rationality in my (1986), specifically in connection with epistemic justification.

can avoid thinking about the possible problem—a classical instance of irrationality.

If we have the ability to do it wrong, what is it that determines when we are doing it right? That is, what makes rational cognition rational? It is a striking fact about human beings that we often find it easy to detect irrational cognitive behavior. How do we do that? Philosophers sometimes speak vaguely of their "philosophical intuitions", but that is to do no more than label the ability. When we catch an agent in irrationality, we know how to perform the task at hand. Knowing how to perform a task consists of knowing what to do as the task unfolds. We detect irrationality by knowing what to do and seeing that the agent does something different.

Knowing how to do something constitutes having *procedural knowledge.* I have many kinds of procedural knowledge. I know how to ride a bicycle, how to do long division, how to speak English, and how to engage in various kinds of epistemic and practical cognition. Having procedural knowledge for how to do something does not dictate that I always do it that way. Sometimes I fall off my mountain bike, make mathematical mistakes, speak ungrammatically, and reason incorrectly. When I know how to do something, I have either a learned or a built-in routine for doing it, but I do not always do it in that way. An important fact about human beings is that we have some ability to detect cases in which we do not conform to our learned or built-in routine. No doubt the functional explanation for this ability is that it enables us to correct our behavior and bring it into conformance with the way we know how to do things.

Because we do often try to bring our behavior into conformance with our procedural knowledge of how to do things, we can regard the learned or built-in routines as playing a normative role of sorts. As such, we can describe the routine in terms of a set of *norms*—rules for how to perform the routine. Chomsky (1957) introduced the *competence/performance distinction* as a way of talking about this. A performance theory regarding some activity is a theory of how people actually perform it. A competence theory is a theory about how they perform it when they are conforming to their procedural knowledge of how to perform it. So a competence theory is, in effect, a description of the procedural norms governing the way people have learned to perform the activity (or the built-on procedural norms if they are not learned).

Chomsky was interested in understanding what theories of grammar are about. His suggestion was that theories of grammar are competence theories of certain aspects of linguistic performance. Speakers of a language know how to speak grammatically, but they do not always do so. Because speakers often speak ungrammatically, a theory of grammar cannot be regarded as a performance theory. However, speakers have the ability to detect when they are diverging from their grammatical norms. Linguists assess grammaticality by asking speakers (or themselves) whether they regard particular sentences as grammatical. When they do this, they say that they are appealing to the "linguistic intuitions" of the language user. On Chomsky's account, these linguistic intuitions are just an exercise of speakers'

ability to tell whether they are conforming to their procedural knowledge when they utter a particular sentence.

Chomsky's account of theories of grammar is now generally accepted in linguistics. In my (1986) I suggested an analogous account of epistemological theories. We know how to perform various cognitive tasks, among them being various kinds of epistemic cognition. Let us take *epistemic norms* to be the norms describing this procedural knowledge. Having this procedural knowledge carries with it the ability to detect when we are not conforming to our procedural norms. My suggestion is that the best way of understanding our epistemological intuitions is to take them as analogous to linguistic intuitions. That is, they are just a reflection of our ability to detect divergences from our epistemic norms. So an epistemological theory is a competence theory of epistemic cognition.

I propose that we extend this account to rationality in general. That is, a theory of rational cognition is a competence theory of human cognition. It describes our norms for how to cognize. I presume that our basic knowledge of how to cognize is built-in rather than learned. It is hard to see how we could learn it without already being able to cognize. So the basic norms for rational cognition are descriptive of our built-in cognitive architecture. More specifically, they are descriptive of those aspects of our cognitive architecture that guide our cognitive performance without rigidly determining it. They are descriptive of the norms provided by our cognitive architecture for how to perform those cognitive tasks over which we have deliberate control.

My reason for adopting this view of human rationality is that it seems to be the best way of making sense of the kind of support that philosophers typically offer for their claims about rationality. They appeal to their philosophical intuitions, but those are utterly mysterious unless we identify them with the familiar ability to monitor our own conformance to our procedural norms.

To summarize my conclusions so far, a competence/performance distinction arises for an agent whose cognitive architecture imposes rules for correct cognition but also enables the agent to violate them. A competence theory is a theory about performances that conform to the rules, and a performance theory is a general theory describing the agent's performance both when it does and when it does not conform to the rules for correct cognition. One way to think of a theory of rationality is as a theory of how to perform cognitive tasks "correctly", that is, in terms of the built-in rules of the cognitive architecture. This is to identify the theory of rationality with a competence theory of cognition. I will use the term "human rationality" to refer to a competence theory of human cognition. This approach generates a concept of rationality that is tightly tied to the details of the human cognitive architecture.

Although I take the preceding to be descriptive of standard philosophical methodology in investigating rationality, it is often useful to take a wider view of rationality, approaching it from the "design stance". We can ask how one might build a cognitive agent that is capable of satisfying various design goals. This immediately raises the thorny issue of what we should

take to be the design goals of human cognition. But it turns out that by approaching cognition from the design stance we can explain many of the more general features of human cognition without saying anything precise about the design goals. For example, for a very wide range of design goals an agent will work better if it is capable of defeasible reasoning, if it treats perceptual input defeasibly, if it is able to reason inductively, if it is able to engage in long range planning, and so on. This generates a "generic" concept of rationality in which we are interested in how cognition might work in cognitive architectures aimed at a broad range of design goals. From this perspective, fine details of human cognition can often be viewed as fairly arbitrary choices in designing a system of cognition. For example, in building an agent, we may want to equip it with a set of rules sufficient for reasoning in the propositional calculus. There are many different sets of inference rules that will suffice for this purpose, and there may be little reason to choose some over others. Thus an arbitrary decision must be made. There is considerable psychological evidence to indicate that modus tollens is not among the built-in inference rules in human beings—it must be learned (Wason 1996; Cheng and Holyoak 1985). Thus from the perspective of human rationality, reasoning in accordance with modus tollens (before learning it through experience) is irrational. But there would be nothing wrong with building a cognitive agent in which modus tollens is built in. Relative to that agent's cognitive architecture, reasoning in accordance with modus tollens prior to learning about it from experience is perfectly rational.

I am primarily interested in understanding rational decision making in human beings. This makes it relative to the human cognitive architecture. However, those details of human cognition that could easily have been otherwise are of less interest to me than those that could not have been changed without adopting a radically different architecture. Thus in studying rational decision making we can ask two kinds of questions. We can ask how a human being should go about making a decision given the cognitive architecture that nature has endowed him with. But we can also evaluate the architecture itself, asking whether it could be significantly improved in various ways without radically altering the general form of the architecture. This second kind of question can in turn have implications for the first kind of question, because as noted above, although we cannot alter our built-in architecture, we often have the ability to employ learned behaviors to override built-in behaviors. Thus if there are better ways to solve decision problems than those dictated by our built-in procedures, we may be able to employ them. Our built-in procedures are often just default procedures, to be employed until we find something better, and when we do find better procedures our built-in architecture itself dictates using them to override the default procedures.

As we proceed, it will be very useful to keep in mind the distinction between evaluating a decision and evaluating a cognitive architecture. Theories of ideal rationality that cannot plausibly be adopted as theories about how real agents ought to make decisions may nevertheless be relevant to the evaluation of cognitive architectures. I will also argue, in chapter 3, that

in at least one respect, human "evaluative cognition" is based upon a rather crude solution to the design problems it aims to solve. We cannot say that humans are irrational for working in the way they do. They cannot help the way they are built. But it is interesting to ask whether we could build a better agent. When the time comes, I will raise this issue again.

4. Decision Making

Before beginning the investigation of how decisions rationally ought to be made, it will be useful to reflect on what goes on in actual decision making. In a particularly simple case, I may just be deciding whether to perform some action *A*. For instance, I may be deciding whether to order the south-western quiche for lunch. This often involves comparing *A* to a number of other alternatives. For instance, should I instead order the chicken salad sandwich? In a particularly simple case, my choice could just be between *A*-ing and not *A*-ing.

It is important to realize that decisions are always made in advance of acting. You cannot literally decide to do something *now*. If by "now" you mean "at this very moment", then either you are already performing the action or you are not performing the action. It is too late to decide. Of course, your decision might be about what to do within the next second or two. But we often have to make decisions far in advance of the time they are to be carried out. This is for at least three reasons. First, I may have to do other things before I can carry out a decision. For instance, if I decide to paint my bedroom, I may have to buy the paint. Second, decisions can involve a whole course of actions rather than a single action. I may decide to paint two rooms, doing the bedroom last. The decision has to be made early enough that I can paint the first room before painting the bedroom, and hence the painting of the bedroom may not occur until some time after the decision is made. When we decide to perform a whole sequence of actions, we are adopting a *plan*. I will say much more about plans over the course of this book.

The third reason decision making often precedes acting by an extended period of time is that decision making can be difficult, consuming considerable cognitive resources and taking quite a bit of time. In the course of making a decision we may have to acquire additional information and we may have to think long and hard about it. We may not have time do all this just shortly before the time to act. We may have to do it well in advance. This is particularly common when we have to perform a number of actions in quick succession. Consider planning a driving route through an unfamiliar city on busy highways. You must plan ahead, memorize where you will go at each intersection, and then follow your plan without further deliberation.

It is because we are resource-bounded cognitive agents that we must often plan well in advance of acting. Having chosen a plan—made a decision—our default procedure must be to follow it automatically. However, if things do not go as expected, we must be able to reopen deliberation and

reconsider the plan. For instance, if you run into unexpectedly high traffic on your chosen route through the city, you must be able to consider changing your plan.

A further complication is that when we plan ahead we will be subject to varying degrees of ignorance about the conditions under which the plan will be executed. For example, I do not normally plan ahead about which traffic lane to use on a particular leg of my route. I decide that in light of the flow of traffic around me as I drive. Decisions about the details of my plans are often best left until the last minute when I will know more about the circumstances in which I am executing the plan. To accommodate this, plans are typically somewhat schematic.[6] I adopt a skeletal plan far in advance, and then slowly fill in the details by making further decisions as the time for acting draws nearer. The end result of such temporally extended planning is a better plan. The resulting plan is not the result of a single act of decision making. It results from a temporally extended sequence of decisions.

There is another reason for adopting skeletal plans. I may have to make a decision before I have time to work out all the details. For example, if I am invited to give a talk at a conference in Copenhagen nine months hence, I may have to decide quickly whether to do it, without having the time to plan exactly what flights I will take. I can work out the details later when I have more time and a lighter cognitive load.

The upshot is that our decisions result in our adopting plans that are schematic to varying degrees. As the time to act draws nearer, planning continues, filling in more details. Note that I may start executing the first part of my plan before filling in the details of the later parts. For instance, I start driving through the city before I decide what traffic lanes to use when I am on the far side of the city.

It is noteworthy that our plans almost never involve precise specifications of when we are going to perform the actions prescribed by the plans. That is not determined until we actually do it. For example, I might decide to buy milk at a convenience store, so I go to the store, take a carton of milk out of the cooler, take it to the cashier, and pay her for it. But I do not decide beforehand at precisely what instant I will hand the money to the cashier. That is not determined until I do it. Furthermore, when I actually do hand the money to the cashier, that does not seem to be a matter of deciding. We certainly do not deliberate about when to do it at the instant we do it. If we were still deliberating, we would not be ready to do it yet, and if we are doing it we must have stopped deliberating at some earlier time. There has to be some point at which deliberation ends and automatic processing takes over. The initiation of the action must be an automatic process rather than one we do deliberately. When we are through deliberating, the action goes on a queue of actions waiting to be performed, and actions are initiated in the order they are retrieved from the queue, without the cognizer thinking about it any further. Philosophers sometimes claim that

[6] This point was emphasized in my (1995).

actions are initiated by a mental act of "willing", but I am not sure what that is supposed to amount to. Do we have to will ourselves to will? If so, we seem to be threatened with an infinite regress. On the other hand, if we can initiate a willing without willing to will, why can't we initiate a finger wiggling without willing that? I think that the philosophers who appeal to willing are over-intellectualizing action initiation. Once I decide to perform the action, and put it on the queue, my action is initiated by my cognitive system, not by me.

One consequence of the schematicity of plans is that I may adopt a plan—form the intention to execute it—but fail to execute it despite the fact that I do not change my mind. For example, I may decide to go to the grocery store this afternoon. But at the time I make this decision I do not decide precisely when to go—just sometime this afternoon. The afternoon passes and I am always busy doing other things, with the result that I do not get to the store. But I did not change my mind about going.

The general picture that emerges from this is that we deliberate and decide on either individual actions or entire plans. Forming an intention amounts to deciding to perform an action or adopt a plan. Adopting a plan has the consequence that, unless we have a lot of cognitive free time, we act on the plan without reconsidering it unless we acquire new information that forces us to reconsider it. On the other hand, we *can* always reconsider our decisions, but as long as we have a pretty good plan, our limited cognitive resources will normally be expended on other more pressing matters.

Although we do not usually reconsider our plans once we have adopted them, execution cannot be entirely automatic because further decision making will be required to fill in the details. It can happen that we are unable to find a good way of filling in the details, in which case the plan will be aborted. For example, when I go to pay for the milk, there may be a power outage that shuts down the cash register, leaving me without a way of paying.

Armed with this understanding of what goes on in decision making, our objective is to investigate what constraints rationality places on the process. How should a cognitive agent, subject to realistic cognitive limitations, decide what to do?

5. Classical Decision Theory and the Optimality Prescription

Throughout contemporary philosophy and cognitive science, one encounters the almost universal presumption that the problem of rational choice is essentially solved by classical decision theory. One of the main conclusions of this book will be that classical decision theory is wrong—when allowed its head, it leads to intuitively incorrect prescriptions about how to act. There is something right about classical decision theory, but the problem of constructing a theory of practical cognition becomes that of replacing classical

decision theory with a more sophisticated theory that retains its insights while avoiding its shortcomings.

The fundamental prescription of classical decision theory is the *optimality prescription*, according to which, when one is deciding what to do, rationality dictates choosing the alternative having the highest expected value. This principle provides the cornerstone on which theories of subjective probability are constructed, underlies so-called belief/desire psychology, and drives most work on practical reasoning. It plays a pervasive role in contemporary philosophy, and has rarely been questioned. But I will argue that the principle is false, for several different, essentially orthogonal, reasons.

By "classical decision theory" I mean the nexus of ideas stemming in part from Ramsey (1926), von Neumann and Morgenstern (1944), Savage (1954), Jeffrey (1965), and others who have generalized and expanded upon it. The different formulations look very different, but the basic prescription of classical decision theory can be stated simply. We assume that our task is to choose an action from a set **A** of *alternative actions*. The actions are to be evaluated in terms of their outcomes. We assume that the *possible outcomes* of performing these actions are partitioned into a set **O** of pairwise exclusive and jointly exhaustive outcomes. In other words, it is logically impossible for two different members of **O** to both occur, and it is logically necessary that some member of **O** will occur. We further assume that we know the probability **PROB**(O/A) of each outcome conditional on the performance of each action. Finally, we assume a *utility measure* $\mathbf{U}(O)$ assigning a numerical "utility value" to each possible outcome. The *expected value* of an action is defined to be a weighted average of the values of the outcomes, discounting each by the probability of that outcome occurring if the action is performed:

$$\mathbf{EV}(A) = \sum_{O \in \mathbf{O}} \mathbf{U}(O) \cdot \mathbf{PROB}(O/A).$$

The crux of classical decision theory is that actions are to be compared in terms of their expected values, and rationality dictates choosing an action that is *optimal*, that is, one such that no alternative has a higher expected value. This is what I am calling "the optimality prescription". To illustrate, suppose we are comparing two actions. We can push button 1, or we can push button 2. If you push button 1, there is then a probability of 1/3 that you will receive \$3, and a probability of 2/3 that you will receive \$6. If you push button 2, there is then a probability of 1/2 that you will receive \$2, and a probability of 1/2 that you will receive \$7. Which button should you push? Computing the expected values:

$$\mathbf{EV}(\text{button 1}) = 3/3 + (6 \times 2)/3 = 5$$

$$\mathbf{EV}(\text{button 2}) = 2/2 + 7/2 = 4.5$$

So the optimality prescription recommends pushing button 1.

Now I turn to some technical details that can be skipped without loss of comprehension. Throughout this book I will isolate such technical material